# Overview of the NTCIR-12 SpokenQuery&Doc-2 Task

Tomoyosi Akiba
Toyohashi University of Technology
1-1 Hibarigaoka, Tohohashi-shi, Aichi, 440-8580, Japan
akiba@cs.tut.ac.jp

Hiromitsu Nishizaki
University of Yamanashi
4-3-11 Takeda, Kofu, Yamanashi, 400-8511, Japan
hnishi@yamanashi.ac.jp

Hiroaki Nanjo
Academic Center for Computing and Media Studies, Kyoto University
nanjo@media.kyoto-u.ac.jp

Gareth J. F. Jones
Dublin City University
Glasnevin, Dublin 9, Ireland
gjones@computing.dcu.ie

## ABSTRACT

This paper presents an overview of the Spoken Query and Spoken Document retrieval (SpokenQuery&Doc-2) task at the NTCIR-12 Workshop. This task included spoken query driven spoken content retrieval (SQ-SCR) and a spoken query driven spoken term detection (SQ-STD) as the two sub-tasks. The paper describes details of each sub-task, the data used, the creation of the speech recognition systems used to create the transcripts, the design of the retrieval test collections, the metrics used to evaluate the sub-tasks and a summary of the results of submissions by the task participants.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## General Terms

Algorithms, Experimentation, Performance

## Keywords

NTCIR-12, spoken document retrieval, spoken queries, spoken content retrieval, spoken term detection

## 1. INTRODUCTION

The NTCIR-12 SpokenQuery&Doc task evaluated information retrieval systems for spoken content retrieval using spoken query input, i.e. speech-driven information retrieval and spoken document retrieval.

Spoken document retrieval (SDR) in the SpokenQuery&Doc task built on the previous NTCIR-11 SpkenQuery&Doc task [4]. It evaluated two SDR tasks: spoken term detection (STD) and spoken content retrieval (SCR). Different from the previous SpokenQuery&Doc task, separate search topics were used for the STD and SCR tasks.

**Spoken Term Detection:** Within spoken documents, find the occurrence positions of a queried term. STD was evaluated based on both efficiency (search time) and effectiveness (precision and recall).

**Spoken Content Retrieval:** In the SCR task, participants were asked to find spoken segments which included relevant information related to a search query, where a segment was either a pre-defined speech segment or a arbitrary length segment. This task was similar to an ad-hoc text retrieval task, except that the target documents are speech data.

The emergence of mobile computing devices means that it is increasingly desirable to interact with computing applications via speech input. The SpokenQuery&Doc task provided the benchmark evaluation using spontaneously spoken queries instead of typed text queries. Here, a spontaneously spoken query means that the query is not carefully arranged before speaking, and is spoken in a natural spontaneous style. Query generated in this way tend to be longer than a typed text query. Note that this spontaneousness contrasts with spoken queries in the form of spoken isolated keywords which are carefully selected in advance, and represent very different situations in terms of speech processing and composition. One of the advantages of such spontaneously spoken queries as input to a retrieval system is that it enables users to easily submit long queries which give systems rich clues for retrieval, although their spontaneous nature means that they are harder to recognise reliably.

Our SQ-SCR tasks were defined not to find whole lecture units, but rather to find shorter relevant speech segments within complete lectures. The search unit is called a slide group segment (SGS). These are naturally defined units based on the speech segment spoken during the display of one or more presentation slides that focus on a single consistent topic within a lecture. The slide-group-segment (SGS) retrieval task required participants to search for relevant SGS units, and was evaluated using a standard mean average precision (MAP) metrics.

The SQ-STD task is almost same as that conducted in the previous NTCIR SpokenDoc task series [1, 2, 3], but is different in that spoken query terms are used instead of text query terms. Different from the previous NTCIR-11 SpokenQuery&Doc-1 SQ-STD task, all the spoken queries for SQ-STD were uttered by speakers in isolation from the SQ-SCR spoken queries. In addition, we conducted a multi-term detection in this SQ-STD task. Each spoken query has one or more terms (keywords). The participants should

find and detect utterance units in which all the terms are included.

The rest of this paper is organized as follows. Sec.2 describes the design and construction of the SpokenQuery&Doc test collection. Sec.3 and Sec.4 describe the task design and the evaluation results of the SQ-SCR sub-task and the SQ-STD sub-task, respectively.

# 2. TEST COLLECTION

In this section, we describe the components of our test collection, including details of the document collection used for the evaluations, construction of the spontaneously spoken query set and transcription of the spoken content.

## 2.1 Document Collection

As well as the NTCIR-11 SpokenQuery&Doc-1 task, **The Corpus of 1st to 7th Spoken Document Processing Workshop (SDPWS1to7)** was used as the document collection for the NTCIR-12 SpokenQuery&Doc-2 task. This was distributed to task participants by the SpokenQuery&Doc task organisers. The corpus consists of the recordings of the first to seventh annual Spoken Document Processing Workshops with slide-change annotation.

Each lecture in the SDPWS1to7 is segmented using pauses that are no shorter than 200 msec. For this purposes of our task, we define each segment to be an Inter-Pausal Unit (IPU). An IPU is short enough to be used to indicate a position in the lecture. Therefore, IPUs are used as the basic unit to be searched in both the STD and SCR tasks. Furthermore, the time points when a lecture presenter transits her/his presentation slides forward are annotated in the SDPWS1to7. This enables us to divide a lecture into a sequence of speech segments each of which is aligned to a single presentation slide, referred to as *a slide segment*.

Generally, a slide segment can be considered to be a semantically consistent unit with a topic related to its corresponding presentation slide. Actually, most single slides individually correspond to a semantic topic. However, sometimes a single topic is found to be covered by a series of slides for some technical reason. For example, one may use a series of slides to give an animation effect. In order to deal with such irregularities, we have grouped a series of contiguous slides into *a slide group*, which corresponds to a single presentation topic as a whole. Note that most slide groups in the collection consist of just a single slide, while the other (a few) groups consist of multiple slides. We refer to a speech segment aligned to a slide group as *a slide group segment*. In the SQ-SCR task in the SpokenQuery&Doc-2, we regard a slide group segment as a search unit, i.e. a document, for retrieval. Therefore, the SCR task is defined as needing to find a set of slide-group-segments that are relevant to a given search topic.

### 2.1.1 Component Files

The component files of the document collection are grouped into two categories; those provided for each lecture and those provided for each IPU. The former are named using the lecture ID, while the latter are named using its IPU ID, which is the lecture ID followed by a sequential number (starting with 0) for each the IPU connected with a hyphen. Each file has its own extension.

We also refer to slide IDs, which are denoted within some of the files. A slide ID is a number series (starting with 1) of the presentation slides.

**VAD file** Voice activity detection (VAD) is first applied on an audio file in order to segment it into a sequence of IPUs. The VAD file records the result of the VAD applied on the audio data of the lecture. Its extension is `.seg`. This records the time stamp of each IPU from the beginning of the lecture, which can then be looked up as necessary.

Each line of a file, which corresponds to an IPU, has two integers formatted as follows:

<start time> <end time>

A unit of the numbers is 1/16000 second from the beginning of the lecture, i.e. 16000 means one second from the beginning.

**Slide group file** This describes the slide groups of the lecture. Its extension is `.grp`. Each line of a file corresponds to a slide group, which is described as a sequence of contiguous slide IDs. Note that, in this file slide IDs are never omitted so that each slide ID appears exactly once in a file.

**Time stamps of slide transitions** This records the time stamp of the start of each presentation slide. Its extension is `.tmg`. Each line is formatted as follows:

<slide ID> [<minutes> ":"] <second>

The second column denotes the start time of a slide from the beginning of a lecture. Note that the first slide of each slide group must have a corresponding line, but the others are not always a line in this file, i.e. some inner slides in a slide group can be omitted.

Notice that, for most of the lectures in the collection time stamps are recorded at second-level granularity, so that they are not accurate enough to locate the exact position in its corresponding audio file. (This limitation arises from the use of off-the-shell software designed for recording of oral presentations, which was used in most of our recordings.)

**Slide-to-IPU alignment file** This describes alignments between the starting time of a slide and an IPU. Its extension is `.align`. Each line is formatted as follows:

<slide ID> <IPU ID> [ "+" ]

A line without "+"' at its end means that the slide denoted by <slide ID> starts at the beginning of the IPU denoted by <IPU ID>, while a slide with "+" at its end means that the slide starts somewhere within the IPU. This file provides an easy way to divide a transcript of a lecture into a set of documents.

**Manual transcription file** This contains a transcript of a lecture created by a human transcriber. Its extention is `.txt`. Each line is formatted as follows.

<IPU ID> ":" <text>

Several tags, which are explained in another document (the annotation manual), are introduced to describe nonverbal events in the text transcript. Among them, the `(s <slide ID>)` tag is used to indicate the position where the slide denoted by <slide ID> is shown for the first time in the lecture.

**Reference automatic transcription** The organizers prepared seven types of automatic transcriptions using two Large Vocabulary Continuous Speech Recognition (LVCSR) decorders, Julius and KALDI.

Three of them, whose file extension is "`_word.jout`", are word-based transcripts made using a Julius decoder with a GMM-HMM-based acoustic model and a word-based trigram model. Other files whose file extension is "`_syll.jout`", are subword-based transcripts made using the Julius decoder with a GMM-HMM-based acoustic model and a syllable-based trigram model. The other differences are in the language models and the acoustic models used in the Julius decoder.

In addition, the organizers prepared two additional types of transcriptions using a KALDI decoder with a DNN-HMM-based acoustic model for this NTCIR-12 evaluation. Their file extension is "`.kout`". The DNN-based KALDI system is almost the same as that described at this web page [1].

The seven automatic transcriptions are referred to using following identifiers:

**REF-WORD-MATCH, REF-SYLLABLE-MATCH** Their file extension is `.matchLM_{word,syll}.jout`. The acoustic model and the language model are trained by using the Corpus of Spontaneous Japanese. (the same as "matched" transcriptions used in the NTCIR-11 SpokenQuery&Doc-1)

**REF-WORD-UNMATCH-LM, REF-SYLLABLE-UNMATCH-LM** Their file extension is `.unmatchLM_{word,syll}.jout`. The acoustic model is trained by using CSJ, while the language model is trained using newspaper articles. (the same as "unmatched" transcriptions used in the NTCIR-11 SpokenQuery&Doc-1)

**REF-WORD-UNMATCH-AMLM** Its file extension is `.unmatchAMLM_word.jout`. Both the acoustic model and the language model are trained using the "unmatched" condition. They are those distributed as Julius dictation kit v4.3.1 [1], whose acoustic and language models are trained by the ASJ Continuous Speech Corpus (JNAS) and the Balanced Corpus of Contemporary Written Japanese (BCCWJ), respectively.

**K-REF-WORD-MATCH** Its file extension is `.word.kout`. The acoustic model and the language model are trained by using the Corpus of Spontaneous Japanese. The DNN-HMM-based acoustic model is trained with the 947 lecture speeches of CSJ, according to the CSJ recipe opened on the GitHub site[2]. The word-based trigram language model and the lexicon are the same as the one used in "REF-WORD-MATCH". Note that the `.kout` file is made of the KALDI-formatted lattice files using the KALDI toolkit. Please see the README file in the test collection if you use this transcription.

[1] `http://kaldi.sourceforge.net/dnn1.html`

[2] `https://github.com/kaldi-asr/kaldi/tree/master/egs/csj`

*(F えーっと)(D と)* 音声認識とかした<息>場合の *(F ま)(F えー)* 場合だとそのテキストが *(F ま)* 話し言葉そのままになるんですけどそれが<息>*(F ま)* 書き言葉の *(D ば)* ものとは *(F ま)(D か)* 書き言葉のものは *(F まー)(F ま)(A ウェブ;Web)* からとってきたりとか *(F ま)* 論文のものだったりとか<息>*(D と)* そういったものは *(F まー)*<息>書き言葉になるんですけどそれとはだいぶ<H>*(D かた)(D き)(D き)* 形式が違うというか<息>そのままではあまり一致しないということなので<息>*(F ま)* それを上手く分ける必要があると思うんですけど<息>*(D と)* その書き言葉と話し言葉を上手く分類というかそれを区別する方法<息>についての説明が知りたいです<息>*(F えー)(D と)(F まー)* どういった特徴量使っているとか *(F まー)(D ど)* どういった手法を使っているとかそういうことをですね

**Figure 1: An example of a query topic (SpokenQueryDoc-SQSCR-formal-0016).**

**K-REF-SYLLABLE-MATCH** Its file extension is `.syll.kout`. The acoustic model is the same as the one for "K-REF-WORD-MATCH". The syllable-based 4-gram language model is used for making syllable-based transcriptions. This language model is trained with 2,525 syllable-based transcriptions of CSJ, which is slightly different from the syllable-based language model used in the Julius decoder.

**Audio file** The audio files of lectures are stored in WAV format for each IPU. The file names are formatted as follows:

<Lecture ID>_<IPU ID>.wav

## 2.2 Query Construction

### 2.2.1 *Collecting Spontaneously Spoken Query Topics*

In order to construct spontaneously spoken query topics for use in the SQ-SCR task, subjective experiments were carried out. Before recording spoken query topics, subjects were asked to look over the proceedings of SDPWS1to7, to select papers they were interested in, and, for each paper, to invent a search topic based on its content described within a paragraph. The selected paragraph was preserved for use later in relevance judgment for topic.

In the recording session, subjects were asked to speak their search topics and their speech was recorded using a close microphone and an IC recorder. Throughout the session, they were not allowed to see their selected paper or any other written material. Therefore, we sought to make the subjects try to recall their search topic by themselves. There was no limitation in speaking time; they could even be silent for a while in order to recall what to say and in order to arrange how to say it. Finally, the session was closed when they felt that they had described their search topic as much as they wished to.

We employed 22 students (21 graduate and 1 undergraduate, 2 females and 20 males) for the experiment. For each

subject, two or four query topics were recorded through our experiment described above, which resulted in a total of 80 topics.

The collected topics were transcribed manually by their creators by themselves. They were also automatically transcribed by using the same ASRs as were used for transcribing the document collection. This resulted in seven types of automatic transcriptions for each spoken topic.

#### 2.2.2 *Collecting Spontaneously Spoken Query Terms*

For the SQ-STD task, we collected spoken queries uttered by 13 subjects (nine males and four females). We prepared two types of queries: one single-term, and the other multi-term. The query type of single-term is the same as the previous test collections in NTCIR-9, 10, and 11. A single-term query is composed of at least one morpheme (word). The definition of morpheme (words) is according to the dictionary ("Unidic" version 1.3.9) used by the morphological analyzer "ChaSen." The query type of multi-term is composed of two or three single-term queries.

The spoken queries are recorded using a headset microphone (close to mouth) and an IC recorder (16 kHz with 16bit sampling). When spoken queries were recorded, we handed a query list to the subjects and gave a few directions to them as follows:

- Speak naturally like using "Google Voice Search."
- For multi-term query, speak multiple words in order-free
- A speaker can insert a breathing pause (short pause) between single-terms. It is okay that a speaker does or does not do this.

Finally, we collected 2,626 utterances (202 kinds of queries *times* 13 subjects) including 767 multi-term query utterances (59 sorts of multi-term queries) and 1,859 single-term query utterances (143 sorts of single-term queries.). We divided these 2,626 query utterances into two groups: one for the dry-run set, and the other for the formal-run set. The query set of the dry-run includes 120 query utterances (40 sorts of queries (10 single- and 40 multi-term) *times* three subjects (two males and one female)). The formal-run query set has 1,620 query utterances (162 sorts of queries (113 single- and 49 multi-term) *times* ten subjects). The queries and the speakers in the dry-run evaluation were not used in the formal-run evaluation. Table 1 shows details of the statistics of the queries, including the number of each query type and their true occurrences in the SDPWS corpus for the SQ-STD query sets.

### 2.3 Relevance Judgment

Relevance judgment for the SQ-SCR slide-group-segment task was performed against slide-group-segments (SGSs) in the document collection based on two clues: the selected paragraph in the paper used by the topic creator (a subject of the experiment described in Sec.2.2.1) to create the topic, and pooling of SGSs submitted by the task participant's systems. The judgment was performed not only on the SGSs specified in their submissions, but also on all the SGSs included in the same candidate lectures.

The assessors were the same as the speakers of the spoken queries. They annotated three levels relevancy:"R" (relevant), "P" (partially relevant), and "I" (irrelevant), on each SGS in their charge, based on both its presentation slide and the manual transcription of its speech segment.

The relevance judgment for the SQ-STD task was obtained automatically by searching for a query term in the manual transcript of the document collection.

**Table 1: The number of query types and their real occurrences (numbers in parentheses) in the SDPWS corpus. The definition of OOV (IV) depends on the ASR dictionary (the matched condition).**

| | | dry-run | formal-run |
|---|---|---|---|
| single | IV | 15 (105) | 63 (573) |
| | OOV | 15 (84) | 50 (352) |
| multi | IV | 5 (15) | 22 (71) |
| | OOV | 5 (10) | 27 (108) |
| Total | IV | 20 (120) | 85 (644) |
| | OOV | 20 (94) | 77 (460) |

## 3. SQ-SCR TASK

### 3.1 Query

For the task data for our evaluation, the organizers provided two set of files: one for spoken queries, and one for text queries. The query topic IDs are given in the names of these files so that the corresponding files can to be used for searching.

#### 3.1.1 *Files for spoken queries*

**Audio file** The audio files of the spoken queries are stored in WAV format. The file names are formatted as follows:

<Query topic ID>.wav

**VAD file** This records the result of the voice activity detection applied on the audio data of the spoken queries. The file names are formatted as follows:

<Query topic ID>.seg

Each line of a file has two integers formatted as follows:

<start time> <end time>

A unit of the numbers is 1/16000 second from the beginning of the query, i.e. 16000 means one second from the beginning.

Note that all the automatic transcripts provided by the task organizers, described below, were obtained by applying ASR on the sequence of the speech segments derived by the VAD process.

**Automatic transcription** This stores an output of a automatic speech recognition of a spoken query. The file names are formatted as follows:

<Query topic ID>_<recognition condition>.jout

The organizers provided seven kinds of recognition results by varying the recognition conditions for each spoken query. The conditions were the same as those used to transcribe the target spoken documents as described in Section 2.1.1.

```
<QUERY-TOPIC-LIST>
  <QUERY id="SpokenQD-SQSCR-dry-0001">
    <TXT file="SpokenQD-SQSCR-dry-0001.txt" />
    <SPK file="SpokenQD-SQSCR-dry-0001.wav">
      <TRANSCRIPTION id="REF-WORD-MATCH"
        file="SpokenQD-SQSCR-dry-001_match_word.jout"
        vad="SpokenQD-SQSCR-dry-001.seg"
        unit="word"
        acoustic-model="match"
        language-model="match" />
      ...
    </SPK>
  </QUERY>
  <QUERY id="SpokenQD-SQSCR-dry-0002">
  ...
  </QUERY>
...
</QUERY-TOPIC-LIST>
```

**Figure 2: An example of a query topic list file.**

### 3.1.2 Files for text queries

**Manual transcription** The manually transcribed text for a spoken query is stored in this file. The file names are formatted as follows:

<Query topic ID>.txt

### 3.1.3 Query Topic List

A query topic list file summarizes the materials described above into a single XML document. It has a single root level tag "<**QUERY-TOPIC-LIST**>". Under the root tag, there are a sequence of tags "<**QUERY**>", each of which corresponds to a single query topic.

A "<**QUERY**>" has one attribute named "id", where its own query topic id is denoted as its value. Within a "<**QUERY**>" tag, three tags named "<**TXT**>", "<**SPK**>", and "<**STD**>" are specified.

- <**TXT**>

  This has one attribute "file" and its value is the file name of the manual transcript of the query topic.

- <**SPK**>

  This has one attribute "file" and its value is the file name of the audio file of the spoken query topic. Under this tag, a set of "<**TRANSCRIPTION**>" tags are described, each of which refers to an automatic transcription of the spoken query. The recognition condition is described in its "id", "vad", "unit", "acoustic-model", and "language-model" attributes. The "id" attribute denotes the identifier of the recognition condition that is same as that used to identify the condition of the target spoken documents. The "vod" attribute denotes the VAD files on which the ASR is applied. The "unit", "acoustic-model", and "language-model" attributes explain the details of the recognition conditions.

Figure 2 shows an example of a query topic list file.

## 3.2 Submission

Each participant was allowed to submit as many search results ("runs") as they wanted. Submitted runs should be prioritized by each group, because a specific number of runs with higher priority would be used for the pooling data for the manual relevance judgments. A priority number should be assigned for each submissions by a participant group, with smaller number having higher priority.

### 3.2.1 File Name

A single run is saved in a single file. Each submission file should have an adequate file name following the next format.

SQSCR-*X*-*T*-*I*-*N*.txt

*X*: System identifier, should be the same as the group ID (e.g., NTC)

*T*: Target task.

- SGS: Slide-Group-Segment retrieval task.
- PAS: Passage retrieval task.

*I*: Input modality.

- SPK: Spoken Query.
- TXT: Text Query.

If a run specifies SPK in this field, it is allowed to use only the query files for spoken queries (Sec.3.1.1) but not the files for text queries (Sec.3.1.2.

*N*: Priority of run (1, 2, 3, ...) for each target document set.

Suppose the group "NTC" submited two files and one file for the slide-group-segment retrieval task by using spoken queries and text queries respectively, and three files for the passage retrieval task by using text queries. Then, the names of the run files should be "SQSCR-NTC-SGS-SPK-1.txt", "SQSCR-NTC-SGS-SPK-2.txt", "SQSCR-NTC-SGS-TXT-1.txt", "SQSCR-NTC-PAS-TXT-1.txt", and "SQSCR-NTC-PAS-TXT-2.txt".

### 3.2.2 Submission Format

The submission files are organized with the following tags. Each file must be a well-formed XML document. It has a single root level tag "<**ROOT**>". Under the root tag, it has three main sections, "<**RUN**>", "<**SYSTEM**>", and "<**RESULT**>".

- <**RUN**>

  <**SUBTASK**> "SQ-SCR" or "SQ-STD". For a SQ-SCR subtask submission, just say "SQ-SCR".

  <**SYSTEM-ID**> System identifier that is the same as the group ID.

  <**PRIORITY**> Priority of the run.

  <**UNIT**> The retrieval unit to be retrieved. "SLIDE-GROUP" if the unit is a slide group as in the slide-group-segment retrieval task. "PASSAGE" if the unit is a passage as in the passage retrieval.

**<TRANSCRIPTION>** The transcription used as the text representation of the target document set. "MANUAL" if it is the manual transcription. "REF-WORD-MATCH", "REF-WORD-UNMATCH-LM", "REF-WORD-UNMATCH-AMLM", "REF-SYLLABLE-MATCH", "REF-SYLLABLE-UNMATCH-LM", "K-REF-WORD-MATCH", "K-REF-SYLLABLE-MATCH", if it is one of the reference automatic transcription provided from the task organizers. "OWN" if it is obtained by a participant's own recognition. "NO" if no textual transcription is used. If multiple transcriptions are used, specify all of them by concatenating with the "," separator.

**<QUERY-TRANSCRIPTION>** The transcription used as the text representation of the spoken queries. "MANUAL" if text queries are used instead of spoken queries. "REF-*" ("*" should be replaced by a transcription Identifier) if one of the reference transcription provided from the task organizers is used. "NO" if no textual transcription is used. If multiple transcriptions are used, specify all of them by concatenating with the "," separator.

- **<SYSTEM>**

  **<OFFLINE-MACHINE-SPEC>**
  **<OFFLINE-TIME>**
  **<INDEX-SIZE>**
  **<ONLINE-MACHINE-SPEC>**
  **<ONLINE-TIME>**
  **<SYSTEM-DESCRIPTION>**

- **<RESULT>**

  **<QUERY>** Each query topic has a single "QUERY" tag with an attribute "id" specified in query topic files (Section 3.1). Within this tag, a list of the following "CANDIDATE" tags is described.

  **<CANDIDATE>** Each potential candidate of a retrieval result has a single "CANDIDATE" tag with the following attributes. The CANDIDATE tags should, but do not necessary to, be sorted in descending order of likelihood.

  **rank** The rank in the result list. "1" for the most likely candidate, incleased one at a time. Required to be totally ordered in a single "QUERY" tag.

  **lecture** The lecture ID specified in the SDPWS1to7.

  **slide** *Used for the slide-group-segment retrieval task.* The first slide ID in a slide group (i.e., a document) that is retrieved as a candidate. If the slide ID that is not first, i.e. second or later, in a slide group is specified, its CANDIDATE tag is always marked wrong in evaluation.

  **ipu-from** *Used for the passage retrieval task.* The Inter Pausal Unit ID, specified in the CSJ, of the first IPU of the retrieved passage (an IPU sequence).

```
<ROOT>
  <RUN>
    <SUBTASK>SQ-SCR</SUBTASK>
    <SYSTEM-ID>TUT</SYSTEM-ID>
    <UNIT>SLIDE-GROUP</UNIT>
    <PRIORITY>1</PRIORITY>
    <TRANSCRIPTION>REF-WORD-UNMATCHED,
      REF-SYLLABLE-UNMATCHED</TRANSCRIPTION>
    <QUERY-TRANSCRIPTION>REF-SYLLABLE-UNMATCHED
      </QUERY-TRANSCRIPTION>
  </RUN>
  <SYSTEM>
    <OFFLINE-MACHINE-SPEC>Xeon 3GHz dual CPU, 4GB mem.
    </OFFLINE-MACHINE-SPEC>
    <OFFLINE-TIME>18:35:23</OFFLINE-TIME>
    ...
  </SYSTEM>
  <RESULT>
    <QUERY id="SpokenQueryDoc0-dry-001">
      <CANDIDATE rank="1" lecture="10-09" slide="8" />
      <CANDIDATE rank="2" lecture="12-12" slide="3" />
      ...
    </QUERY>
    <QUERY id="SpokenQueryDoc0-dry-002">
      ...
    </QUERY>
  </RESULT>
</ROOT>
```

**Figure 3: An example of a submission file.**

  **ipu-to** *Used for the passage retrieval task.* The Inter Pausal Unit ID, specified in the CSJ, of the last IPU of the retrieved passage (an IPU sequence).

  **NOTE:** The IPU sequences specified in a single "QUERY" tag are required to be *exclusive* each other; i.e. no two intervals in a "QUERY", each of which is specified by "CANDIDATE" tag, are not allowed to have a common IPU.

Figure 3 shows an example of a submission file.

## 3.3 Evaluation Measures

### 3.3.1 Slide-Group-Segment Retrieval

Mean Average Precision (MAP) was used as the official evaluation measure for lecture retrieval For each query topic, the top 1000 documents were evaluated.

Given a question $q$, suppose the ordered list of documents $d_1 d_2 \cdots d_{|D|} \in D_q$ was submitted as the retrieval result. Then, $AveP_q$ is calculated as follows.

$$AveP_q = \frac{1}{|R_q|} \sum_{i=1}^{|D_q|} include(d_i, R_q) \frac{\sum_{j=1}^{i} include(d_j, R_q)}{i} \tag{1}$$

where

$$include(a, A) = \left\{ \begin{array}{ccc} 1 & \cdots & a \in A \\ 0 & \cdots & a \notin A \end{array} \right. \tag{2}$$

Alternatively, given the ordered list of correctly retrieved documents $r_1 r_2 \cdots r_M (M \leq |R_q|)$, $AveP_q$ is calculated as

follows.

$$AveP_q = \frac{1}{|R_q|} \sum_{k=1}^{M} \frac{k}{rank(r_k)} \quad (3)$$

where $rank(r)$ is the rank that the document $r$ is retrieved.

**MAP** is the mean of the $AveP$ over all query topics $Q$.

$$\mathbf{MAP} = \frac{1}{|Q|} \sum_{q \in Q} AveP_q \quad (4)$$

## 3.4 Result

Four groups with a total 33 runs submitted their results for the formal-run of the SQ-SCR slide-group-segment task. The group ID and their submitted runs are listed in Table 2.

### 3.4.1 Baseline

Our baseline runs were implemented by applying conventional methods for IR simply on manual and four word-based reference transcriptions of document collection from, in the same way, manual and four reference transcriptions of spontaneously spoken query. Only nouns were used for both indexing the document collections and extracting keywords from a query, which were processed by applying the Japanese morphological analysis tool. We simply used the off-the-shelf IR system, lucene, for retrieval.

The retrieval results of all combinations of transcripts between query and document in terms of MAP are shown in Table3.

### 3.4.2 Evaluation Results

Table 4 shows the run-by-run evaluation results of the slide-group-segment retrieval task, where the runs are grouped by query transcription and document transcription type used.

# 4. SQ-STD TASK

## 4.1 Query

The query terms used for the SQ-STD task are put together into a single file written in an XML format, called query term list. This has a single root level tag "<**QUERY-TERM-LIST**>". Under the root tag, there are a sequence of tags "<**QUERY**>", each of which corresponds to a single query term.

A "<**QUERY**>" tag has two attributes named "id" and "speaker", in which query ID and speaker ID are denoted as their values, respectively.

- <**TEXT**>

  It is used to describe the materials used for the STD task from text queries. It has at least two attributes "term$N$" and "pron$N$", where $N$ means the sequensal number of term. If a query has two terms, <**TEXT**> has four attirbutes like "term1", "pron1", "term2" and "pron2". The value of the "term$N$" tag is the manually transcribed text of the query term, while that of the "pron$N$" tag is the Japanese pronunciation of the query term written in a Japanese KATAKANA sequence. Note that each term is composed of one or a few morphological words obtained using a morphological analyzer "Chasen-2.4.2" with a morphological dictionary "Unidic-1.3.9". OOV decision for each term is based on morphological words and the lexicon used in the LVCSR decorders.

  Notice that, for the judgment of the term's occurrence in the golden file, "text$N$" is searched against the manual transcriptions, while the "pron$N$" is never considered for the judgment. Furthermore, the organizers do **not** assure the participants of the correctness of what described in the "pron$N$" fields, so the participants should take the responsible for using it. Nevertheless, the organizers believes it should help participants to predict the term's pronunciation.

- <**SPEECH**>

  Under this tag, the materials used for the SQ-STD task from spoken queries are described. It has two attributes "spk-id" and "spoken-query-file". The value of "spk-id" is the speaker's ID, and "spoken-query-file" shows a wave-formatted file of the spoken query uttered by the "spk-id" speaker.

  Participants could can propose STD methods for the query by example search, including acoustic feature level matching method between spoken documents and a spoken query. Of course, they could perform ASR for the spoken queries using their OWN ASR decorder(s). On the other hand, the organizers prepared ASR transcripts of the spoken queries using the Julius and the KALDI decorders. The details are written on the README file in the distributed test collection.

```
<QUERY-TERM-LIST>
  <QUERY id="SpokenQueryDoc2-SQSTD-dry-001">
    <TEXT term1="アーティキュレーション"
        pron1="アーティキュレーション" />
    <SPEECH spk-id="D01"
        spoken-query-file="D01-dry-001.wav" />
    <SPEECH spk-id="D02"
        spoken-query-file="D02-dry-001.wav" />
    <SPEECH spk-id="D03"
        spoken-query-file="D03-dry-001.wav" />
  </QUERY>
  <QUERY id="SpokenQueryDoc2-SQSTD-dry-002">
    <TEXT term1="インターフェース"
        pron1="インターフェース"
        term2="ロボット" pron2="ロボット" />
    <SPEECH spk-id="D01"
        spoken-query-file="D01-dry-002.wav" />
    <SPEECH spk-id="D02"
        spoken-query-file="D02-dry-002.wav" />
    <SPEECH spk-id="D03"
        spoken-query-file="D03-dry-002.wav" />
  </QUERY>
...
</QUERY-TERM-LIST>
```

**Figure 4: An example of a qyery term list file.**

Figure 4 shows an example of a query term list file.

### 4.1.1 Audio files

The audio files of spoken queries are stored in WAV format for each query and speaker. The file names are formatted as follows.

<Speaker ID>-dry-<Query ID[3]>.wav ... for the

[3]Only sequencial number in three digits.

**Table 2: SQ-SCR task participants.**

| Group ID | Group Name, Organization | SGS-SPK | SGS-TXT |
|---|---|---|---|
| AKBL | Akiba Laboratory, Toyohashi University of Technology | 2 | 6 |
| DCU | Dublin City University, APAPT Center | 4 | 8 |
| HYM16 | Laboratorie de professeur Chat Noir Gifu University | 4 | |
| UB | UB, University at Buffalo | | 7 |

**Table 3: MAP values (%) of the baseline results.**

| query \document | MANUAL | K-MATCH | MATCH | UNMATCH-LM | UNMATCH-AMLM |
|---|---|---|---|---|---|
| MANUAL | 13.6 | 13.2 | 11.3 | 6.04 | 7.50 |
| K-MATCH | 13.0 | 13.6 | 12.3 | 5.76 | 6.76 |
| MATCH | 11.1 | 12.0 | 10.5 | 4.31 | 6.03 |
| UNMATCH-LM | 5.13 | 6.04 | 4.25 | 3.98 | 3.07 |
| UNMATCH-AMLM | 8.22 | 7.36 | 6.99 | 3.21 | 4.18 |

dry-run queries

<Speaker ID>-formal-<Query ID[4]>.wav ... for the formal-run queries

## 4.2 Transcriptions for spoken queries

The organizers provided seven types of transcripts generated using the two LVCSR systems. The ASR conditions are identical to those for the transcripts of the spoken documents.

Note that all the transcription files of the spoken queries are recorded in UTF-8 format.

## 4.3 Submission

Each participant was allowed to submit as many search results ("runs") as they wished. Submitted runs should be prioritized by each group. Priority number should be assigned through all submissions of a participant, and smaller number has higher priority.

### 4.3.1 File Name

A single run was saved in a single file. Each submission file should have an adequate file name following the next format.

SQSTD-*X*-*T*-*I*-*N*.txt

*X*: System identifier that is the same as the group ID (e.g., NTC)

*T*: Target task.

- IPU: IPU retrieval task.

For SQ-STD task submission, just say "IPU".

*I*: Input modality.

- SPK: Spoken Query.
- TXT: Text Query.

*N*: Priority of run (1, 2, 3, ...) for each target docuemnt set.

[4]Only sequencial number in three digits.

For example, if the group "NTC" submitted two files and three files by using spoken queries and text queries, respectively, then the names of the run files should be "SQSTD-NTC-IPU-SPK-1.txt", "SQSTD-NTC-IPU-SPK-2.txt", "SQSTD-NTC-IPU-TXT-1.txt", "SQSTD-NTC-IPU-TXT-2.txt", and "SQSTD-NTC-IPU-TXT-3.txt".

### 4.3.2 Submission Format

The submission files are organized with the following tags. Each file must be a well-formed XML document. It has a single root level tag "<**ROOT**>". It has three main sections, "<**RUN**>", "<**SYSTEM**>", and "<**RESULT**>".

- <**RUN**>

  <**SUBTASK**> "SQ-STD" or "SQ-STD". For a SQ-STD subtask submission, just say "SQ-STD".

  <**SYSTEM-ID**> System identifier that is the same as the group ID.

  <**PRIORITY**> Priority of the run.

  <**TRANSCRIPTION**> The transcript used as the text representation of the target document set. "MANUAL" if it is the manual transcription. "REF-WORD-MATCH", "REF-WORD-UNMATCH-LM", "REF-WORD-UNMATCH-AMLM", "REF-SYLLABLE-MATCH", "REF-SYLLABLE-UNMATCH-LM", "K-REF-WORD-MATCH", "K-REF-SYLLABLE-MATCH", if it is one of the reference automatic transcript provided from the task organizers. "OWN" if it is obtained using a participant's own recognition. "NO" if no textual transcription is used. If multiple transcriptions are used, specify all of them by concatenating with the "," separator.

  <**QUERY-TRANSCRIPTION**> The transcript used as the text representation of the spoken queries. "REF-WORD-MATCH", "REF-WORD-UNMATCH-LM", "REF-WORD-UNMATCH-AMLM" "REF-SYLLABLE-MATCH", "REF-SYLLABLE-UNMATCH-LM", "K-REF-WORD-MATCH", "K-REF-SYLLABLE-MATCH", if it is one of the reference automatic

**Table 4: SQ-SCR slide-group-segment retrieval task result (%).**

| run ID | Query Transcription | Document Transcription | MAP |
|---|---|---|---|
| AKBL-SGS-SPK-1 | K-REF-WORD-MATCH | K-REF-WORD-MATCH | 13.6 |
| HYM16-SGS-SPK-1 | K-REF-WORD-MATCH | K-REF-WORD-MATCH | 23.9 |
| HYM16-SGS-SPK-2 | K-REF-WORD-MATCH | K-REF-WORD-MATCH | 24.2 |
| HYM16-SGS-SPK-3 | K-REF-WORD-MATCH | K-REF-WORD-MATCH | 23.1 |
| HYM16-SGS-SPK-4 | K-REF-WORD-MATCH | K-REF-WORD-MATCH | 25.2 |
| UB-SGS-TXT-7 | K-REF-WORD-MATCH | K-REF-WORD-MATCH | 19.5 |
| AKBL-SGS-SPK-1 | REF-WORD-MATCH | REF-WORD-MATCH | 10.5 |
| DCU-SGS-SPK-1 | REF-WROD-MATCH | REF-WORD-MATCH | 24.0 |
| DCU-SGS-SPK-7 | REF-WROD-MATCH | REF-WORD-MATCH | 18.3 |
| DCU-SGS-SPK-8 | REF-WROD-MATCH | REF-WORD-MATCH | 25.2 |
| DCU-SGS-SPK-9 | REF-WROD-MATCH | REF-WORD-MATCH | 18.8 |
| UB-SGS-TXT-2 | REF-WROD-MATCH | REF-WORD-MATCH | 11.3 |
| UB-SGS-TXT-3 | REF-WROD-MATCH | REF-WORD-MATCH | 9.9 |
| UB-SGS-TXT-4 | REF-WROD-MATCH | REF-WORD-MATCH | 11.3 |
| UB-SGS-TXT-5 | REF-WROD-MATCH | REF-WORD-MATCH | 9.7 |
| UB-SGS-TXT-6 | REF-SYLLABLE-MATCH | REF-SYLLABLE-MATCH | 2.5 |
| AKBL-SGS-TXT-2 | MANUAL | REF-WORD-MATCH | 19.6 |
| AKBL-SGS-TXT-3 | MANUAL | REF-WORD-MATCH | 9.0 |
| DCU-SGS-TXT-2 | MANUAL | REF-WORD-MATCH | 23.8 |
| DCU-SGS-TXT-6 | MANUAL | REF-WORD-MATCH | 27.9 |
| DCU-SGS-TXT-10 | MANUAL | REF-WORD-MATCH | 21.2 |
| DCU-SGS-TXT-11 | MANUAL | REF-WORD-MATCH | 21.7 |
| AKBL-SGS-TXT-4 | MANUAL | REF-WORD-UNMATCH-AMLM | 9.1 |
| AKBL-SGS-TXT-5 | MANUAL | REF-SYLLABLE-MATCH | 9.7 |
| AKBL-SGS-TXT-6 | MANUAL | REF-SYLLABLE-UNMATCH-LM | 5.8 |
| AKBL-SGS-TXT-1 | MANUAL | MANUAL | 20.8 |
| DCU-SGS-TXT-1 | MANUAL | MANUAL | 34.2 |
| DCU-SGS-TXT-7 | MANUAL | MANUAL | 34.3 |
| DCU-SGS-TXT-8 | MANUAL | MANUAL | 29.3 |
| DCU-SGS-TXT-9 | MANUAL | MANUAL | 27.8 |
| UB-SGS-TXT-1 | MANUAL | MANUAL | 19.5 |

transcription provided from the task organizers. "OWN" if it is obtained by a participant's own recognition. "NO" if no textual transcription is used. If multiple transcriptions are used, specify all of them by concatenating with the "," separator.

- <**SYSTEM**>

  <**OFFLINE-MACHINE-SPEC**>
  <**OFFLINE-TIME**>
  <**INDEX-SIZE**>
  <**ONLINE-MACHINE-SPEC**>
  <**ONLINE-TIME**>
  <**SYSTEM-DESCRIPTION**>

- <**RESULT**>

  <**QUERY**> Each query term has a single "QUERY" tag with an attribute "id" specified in a query term list (Section 4.1) and an attribute "speaker". The "speaker" attribute indicates which speaker uttered the query. This attribute has "TEXT" value when a text query is used to search. Within this tag, a list of the following "TERM" tags is described.

  <**TERM**> Each potential detection of a query term has a single "TERM" tag with the following attributes.

  **lecture** The searched lecture ID.

  **ipu** The searched Inter Pausal Unit ID.

  **score** The detection score indicating the likelihood of the detection. The greater is more likely.

  **detection** The binary ("YES" or "NO") decision of whether or not the term should be detected to make the optimal evaluation result.

Figure 5 and 6 show examples of submission files for the spoken query set and text query set, respectively.

## 4.4 Evaluation Measures

The evaluation measures for effectiveness of STD are Mean Average Precision (MAP), Term-Weighted Value (TWV) based on Decision Error Treadoff (DET) curve, and F-measure basend on recall-precision curve.

MAP is a rank-based measure calculated the mean value based on the macro averaged precision value for each query. This measure is not considered absolute values of detection scores but ranking of the detections for the search query. It

```
<ROOT>
<RUN>
  <SUBTASK>SQ-STD</SUBTASK>
  <SYSTEM-ID>BASELINE</SYSTEM-ID>
  <PRIORITY>1</PRIORITY>
  <TRANSCRIPTION>REF-WORD-UNMATCH-LM,
    REF-SYLLABLE-UNMATCH-LM</TRANSCRIPTION>
  <QUERY-TRANSCRIPTION>K-REF-WORD-MATCH
    </QUERY-TRANSCRIPTION>
</RUN>
<SYSTEM>
  <OFFLINE-MACHINE-SPEC>Xeon 3GHz dual CPU,
    4GB memory</OFFLINE-MACHINE-SPEC>
  <OFFLINE-TIME>66923</OFFLINE-TIME>
  <INDEX-SIZE>2000</INDEX-SIZE>
  <ONLINE-MACHINE-SPEC>Xeon 3GHz dual CPU,
    4GB memory</ONLINE-MACHINE-SPEC>
  <ONLINE-TIME>15</ONLINE-TIME>
  <SYSTEM-DESCRIPTION>baseline system
    </SYSTEM-DESCRIPTION>
</SYSTEM>
<RESULT>
  <QUERY id="SpokenQueryDoc2-SQSTD-dry-001"
        speaker="D01">
    <TERM lecture="07-01" ipu="0000" score="1.000"
        detection="YES" />
    <TERM lecture="07-01" ipu="0001" score="0.500"
        detection="NO" />
    ...
  </QUERY>
  <QUERY id="SpokenQueryDoc2-SQSTD-dry-002"
        speaker="D01">
    <TERM lecture="09-14" ipu="0111" score="1.000"
        detection="YES" />
    <TERM lecture="11-05" ipu="0212" score="0.650"
        detection="NO" />
    ...
  </QUERY>
  <QUERY id="SpokenQueryDoc2-SQSTD-dry-003"
        speaker="D01">
    <TERM lecture="13-01" ipu="0001" score="1.000"
        detection="YES" />
    <TERM lecture="10-18" ipu="0022" score="0.995"
        detection="YES" />
    ...
  </QUERY>
  ...
  <QUERY id="SpokenQueryDoc2-SQSTD-dry-001"
        speaker="D02">
    <TERM lecture="07-01" ipu="0000" score="1.000"
        detection="YES" />
    <TERM lecture="07-01" ipu="0001" score="0.500"
        detection="NO" />
    ...
  </QUERY>
  <QUERY id="SpokenQueryDoc2-SQSTD-dry-002"
        speaker="D02">
    <TERM lecture="09-14" ipu="0111" score="1.000"
        detection="YES" />
    <TERM lecture="11-05" ipu="0212" score="0.650"
        detection="NO" />
    ...
  </QUERY>
  ...
  <QUERY id="SpokenQueryDoc2-SQSTD-dry-001"
        speaker="D03">
  ...
</RESULT>
</ROOT>
```

**Figure 5: An example of a submission file for the spoken query set.**

```
<ROOT>
<RUN>
  <SUBTASK>SQ-STD</SUBTASK>
  <SYSTEM-ID>BASELINE</SYSTEM-ID>
  <PRIORITY>1</PRIORITY>
  <TRANSCRIPTION>K-REF-WORD-MATCH,
    K-REF-SYLLABLE-MATCH</TRANSCRIPTION>
  <QUERY-TRANSCRIPTION>MANUAL
    </QUERY-TRANSCRIPTION>
</RUN>
<SYSTEM>
  <OFFLINE-MACHINE-SPEC>Xeon 3GHz dual CPU,
    4GB memory</OFFLINE-MACHINE-SPEC>
  <OFFLINE-TIME>66923</OFFLINE-TIME>
  <INDEX-SIZE>2000</INDEX-SIZE>
  <ONLINE-MACHINE-SPEC>Xeon 3GHz dual CPU,
    4GB memory</ONLINE-MACHINE-SPEC>
  <ONLINE-TIME>15</ONLINE-TIME>
  <SYSTEM-DESCRIPTION>baseline system
    </SYSTEM-DESCRIPTION>
</SYSTEM>
<RESULT>
  <QUERY id="SpokenQueryDoc2-SQSTD-dry-001"
        speaker="TEXT">
    <TERM lecture="12-08" ipu="0100" score="1.000"
        detection="YES" />
    <TERM lecture="12-08" ipu="0009" score="0.750"
        detection="NO" />
    ...
  </QUERY>
  <QUERY id="SpokenQueryDoc2-SQSTD-dry-002"
        speaker="TEXT">
    <TERM lecture="09-01" ipu="0015" score="1.000"
        detection="YES" />
    <TERM lecture="10-13" ipu="0232" score="0.995"
        detection="YES" />
    ...
  </QUERY>
  ...
</RESULT>
</ROOT>
```

**Figure 6: An example of a submission file for the text query set.**

can be calculated as follows:

$$MAP = \frac{1}{Q}\sum_{i=1}^{Q} AveP(i) \quad (5)$$

where Q is the number of queries and $AveP(i)$ means the average precision of the $i$-th query of the query set. The average precision is calculated by averaging of the precision values computed at the point of each of the relevant terms in the list in which retrieved terms are ranked by a relevance measure.

$$AveP(i) = \frac{1}{Rel_i}\sum_{r=1}^{N_i}(\delta_r \cdot Precision_i(r)) \quad (6)$$

where $r$ is the rank, $N_i$ is the rank number at which the all relevance terms of query $i$ are found, and $Rel_i$ is the number of the relevance terms of query $i$. $\delta_r$ is a binary function on the relevance of a given rank $r$.

TWV is widely used to evaluate STD performances in some kinds of STD test collection such as the NIST KWS evaluations. TWV at the decision point specified by a participant's search system is called Actual TWV (ATWV), and TWV at the maximum decision point on the DET curve is also called Maximum TWV (MTWV). A DET curve is drawn by using miss and false alarm probabilities which are calculated from the detections whose scores are equal to a threshold value or more than a threshold. These probabilities are macro-averaged over all the queries. The calculation method of TWV is shown in the evaluation plan document [5] on the NIST STD 2006 evaluation web page. TWV is more sensitive to false alarms (incorrectly detected terms) compared to MAP.

Finally, F-measures at both the decision point specified by a participant's system and the maximum decision point on a recall-precision curve are used for evaluation. Recall and precision values are obtained by micro- and macro-averaging over all the queries on the test collection. These measures were used in the previous STD evaluations of NTCIR-9, NTCIR-10 and NTCIR-11. A micro-average-based recall-precision curve cannot evaluate each query, but only a whole query set. Therefore, a search system is needed to search all sorts of terms accurately.

## 4.5 Results

### 4.5.1 Baseline

We used the typical Kaldi KWS system[6] [5] as the baseline system for the evaluation on the SQ-STD task. The Kaldi KWS system consists of two phases: the ASR system based on the Kaldi ASR toolkit decodes the SDPWS corpus (search collection) and generates corresponding recognition lattices and the KWS module that makes the index for the lattices and searches a keyword from the generated index.

In the baseline, we performed automatic syllable recognition of the SDPWS corpus for making the index for search using the Kaldi toolkit because we deal with OOV queries. It is well-known that a subword-based transcription is robust for an OOV query [1, 6]. The syllable-based recognition system is the same as "K-REF-SYLLABLE-MATCH" described in Sec. 2.

In the search phase, a text-formed query is converted into a syllable sequence based on the pronunciation of the query, and furthermore, it also is transformed to a finite state transducer. If a query is a multi-term query (consist of two or more single-terms), each single-term is separately searched for the index. We can get the final STD score for the multi-term query by averaging each score of the single-terms.

On the other hand, we did not make the number of terms contained in a query known to the participants. Because of this, participants systems cannot learn the number of varieties of terms in advance. In the case of the baseline system, terms are automatically segmented depending on short pause information by the ASR system. Each automatically segmented-term is separately searched for the index, and each result is fashioned by the same method as for the text-formed query search.

### 4.5.2 Evaluation Results

Three teams participated in the STD subtask; two teams submitted results for the text query set and all teams submitted the results for the spoken query set.

Table 5 and Table 7 show the evaluation results of all the systems submitted by the participants, including the baseline system described in Sect. 4.5.1. Table 5 and Table 7 summarize the STD performances on the text query set and the spoken query set respectively. These tables include actual F-measure, which is calculated using the recall and the precision values when an STD system decides the optimal threshold whether a detection candidate is outputted or not. They also include maximum F-measure that is the optimal point on the recall-precision curve. These actual and maximum F-measures are micro- and macro-averaged on the test collection. The remaining MAP, ATWV, and MTWV are macro-averaged for all the queries in the test collection.

Table 6 and Table 8 also conclude the system descriptions, in which the types of transcripts and computer environments used in each STD system are described. In addition, the processing times of each STD system is also shown. Team "ALPS" used their own transcripts (OWN), based on the Kaldi ASR system.

# 5. CONCLUSIONS

This paper introduced the overview of the Spoken Query and Spoken Document Retrieval Task (SpokenQuery&Doc-2) in NTCIR-12 Workshop.

# Acknowledgement

This work was supported by KAKENHI 25330130, 26282049, and 15K00254.

# 6. REFERENCES

[1] T. Akiba et al. Overview of the IR for spoken documents task in NTCIR-9 workshop. In *Proceedings of the Ninth NTCIR Workshop Meeting*, pages 223–235, 2011.

[2] T. Akiba et al. Designing an evaluation framework for spoken term detection and spoken document retrieval at the NTCIR-9 SpokenDoc task. In *Proceedings of International Conference on Language Resources and Evaluation*, 2012.

[5] http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf

[6] http://kaldi.sourceforge.net/kws.html

**Table 5: Summary on the STD performances for the text query set.**

| system ID | micro ave. | | macro ave. | | | | |
|---|---|---|---|---|---|---|---|
| | Actual F. | Max. F. | Actual F. | Max. F. | MAP | ATWV | MTWV |
| BL-1 | 0.6602 | 0.6803 | 0.7372 | 0.7372 | 0.7054 | 0.6660 | 0.6660 |
| ALPS-1 | 0.5793 | 0.6054 | 0.5965 | 0.6646 | 0.7445 | 0.4641 | 0.6420 |
| ALPS-2 | 0.7099 | 0.7099 | 0.7861 | 0.7873 | 0.8401 | 0.7251 | 0.7931 |
| ALPS-3 | 0.7188 | 0.7289 | 0.8258 | 0.8258 | 0.8655 | 0.7989 | 0.8242 |
| ALPS-4 | 0.2780 | 0.5462 | 0.5582 | 0.6451 | 0.7224 | 0.3491 | 0.5108 |
| ALPS-5 | 0.5508 | 0.5798 | 0.5645 | 0.6209 | 0.6865 | 0.4297 | 0.4962 |
| ALPS-6 | 0.4399 | 0.4676 | 0.4367 | 0.5777 | 0.5164 | 0.3187 | 0.4434 |
| IWAPU-1 | 0.6229 | 0.7269 | 0.7999 | 0.8132 | 0.8555 | 0.7891 | 0.7915 |
| IWAPU-2 | 0.6577 | 0.7308 | 0.7976 | 0.8045 | 0.8404 | 0.7921 | 0.7977 |
| IWAPU-3 | 0.6004 | 0.7040 | 0.7935 | 0.8004 | 0.8398 | 0.7622 | 0.7656 |
| IWAPU-4 | 0.5988 | 0.6967 | 0.7769 | 0.7854 | 0.8233 | 0.7467 | 0.7470 |
| IWAPU-5 | 0.4827 | 0.6529 | 0.7494 | 0.7786 | 0.8083 | 0.7399 | 0.7503 |
| IWAPU-6 | 0.4802 | 0.6530 | 0.7412 | 0.7745 | 0.8018 | 0.7293 | 0.7406 |
| IWAPU-7 | 0.6865 | 0.6963 | 0.7509 | 0.7580 | 0.8069 | 0.6831 | 0.7375 |
| IWAPU-8 | 0.4017 | 0.5124 | 0.6506 | 0.6525 | 0.6080 | 0.5426 | 0.5461 |
| IWAPU-9 | 0.3652 | 0.4780 | 0.6359 | 0.6390 | 0.5695 | 0.5020 | 0.5209 |
| IWAPU-10 | 0.6027 | 0.6239 | 0.6851 | 0.7026 | 0.7253 | 0.5824 | 0.6370 |
| IWAPU-11 | 0.5583 | 0.5635 | 0.6536 | 0.6652 | 0.6989 | 0.5468 | 0.6009 |
| IWAPU-12 | 0.6664 | 0.7499 | 0.7849 | 0.8159 | 0.8641 | 0.8085 | 0.8085 |

**Table 6: Summary on the system descriptions for the text query set.**

| system ID | transcription type (target) | transcription type (query) | offline time [s] | index size [MB] | machine SPEC | online time [s] |
|---|---|---|---|---|---|---|
| BL-1 | K-REF-SYLLABLE-MATCH | MANUAL | 100 | 256 | Core i7 3720QM 2.6GHz, 16GB, no GPU | 30 |
| ALPS-1 | OWN | MANUAL | 3221 | 698 | Xeon E5-2630v3 2.4GHz, 128GB, GTX TitanX | 2188 |
| ALPS-2 | OWN | MANUAL | 3221 | 698 | Xeon E5-2630v3 2.4GHz, 128GB, GTX TitanX | 1543 |
| ALPS-3 | OWN | MANUAL | 223 | 630 | Xeon E5-2630v3 2.4GHz, 128GB, GTX TitanX | 1506 |
| ALPS-4 | OWN | MANUAL | 3221 | 698 | Xeon E5-2630v3 2.4GHz, 128GB, GTX TitanX | 2183 |
| ALPS-5 | OWN | MANUAL | 3221 | 698 | Xeon E5-2630v3 2.4GHz, 128GB, GTX TitanX | 1814 |
| ALPS-6 | OWN | MANUAL | 3221 | 68 | Xeon E5-2630v3 2.4GHz, 128GB, GTX TitanX | 77 |
| IWAPU-1 | K-REF-WORD-MATCH | MANUAL | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 1147 |
| IWAPU-2 | K-REF-SYLLABLE-MATCH | MANUAL | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 1165 |
| IWAPU-3 | K-REF-WORD-MATCH | MANUAL | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 1128 |
| IWAPU-4 | K-REF-SYLLABLE-MATCH | MANUAL | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 1126 |
| IWAPU-5 | K-REF-WORD-MATCH | MANUAL | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 1120 |
| IWAPU-6 | K-REF-SYLLABLE-MATCH | MANUAL | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 1119 |
| IWAPU-7 | K-REF-WORD-MATCH | MANUAL | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 50 |
| IWAPU-8 | K-REF-WORD-MATCH | MANUAL | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 29 |
| IWAPU-9 | K-REF-SYLLABLE-MATCH | MANUAL | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 31 |
| IWAPU-10 | OWN | MANUAL | 30 | 50 | Core i7 4770 3.5GHz 16GB, GTX 750Ti | 36 |
| IWAPU-11 | OWN | MANUAL | 30 | 50 | Core i7 4770 3.5GHz 16GB, GTX 750Ti | 45 |
| IWAPU-12 | K-REF-WORD-MATCH | MANUAL | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 1198 |

[3] T. Akiba et al. Overview of the NTCIR-10 SpokenDoc-2 task. In *Proceedings of the 10th NTCIR Conference*, pages 573–587, 2013.

[4] T. Akiba, H. Nishizaki, H. Nanjo, and G. J. F. Jones. Overview of the ntcir-11 spokenquery&doc task. In *Proceedings of the 11th NTCIR Conference*, pages 350–364, 2014.

[5] G. Chen, S. Khudanpur, D. Povey, J. Trmal, D. Yarowsky, and O. Yilmaz. Quantifying the value of pronunciation lexicons for keyword search in lowresource languages. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8560–8564, May 2013.

[6] H. Nishizaki, T. Akiba, K. Aikawa, T. Kawahara, and T. Matsui. Evaluation Framework Design of Spoken Term Detection Study at the NTCIR-9 IR for Spoken Documents Task. *Journal of Natural Language Processing*, 19(4):329–350, 2012.

**Table 7: Summary on the STD performances for the spoken query set.**

| system ID | micro ave. | | macro ave. | | | | |
|---|---|---|---|---|---|---|---|
| | Actual F. | Max. F. | Actual F. | Max. F. | MAP | ATWV | MTWV |
| BL-1 | 0.2776 | 0.4709 | 0.5589 | 0.5711 | 0.5413 | 0.3731 | 0.4686 |
| ALPS-1 | 0.4939 | 0.5095 | 0.4332 | 0.5841 | 0.6390 | 0.3371 | 0.5143 |
| ALPS-2 | 0.0942 | 0.5606 | 0.0362 | 0.6045 | 0.7362 | 0.0291 | 0.5344 |
| ALPS-3 | 0.0018 | 0.5256 | 0.0011 | 0.5261 | 0.7756 | 0.0006 | 0.4743 |
| ALPS-4 | 0.0108 | 0.1828 | 0.0046 | 0.4942 | 0.5221 | 0.0037 | 0.1685 |
| ALPS-5 | 0.4801 | 0.4983 | 0.3712 | 0.5232 | 0.5714 | 0.2963 | 0.4360 |
| ALPS-6 | 0.3636 | 0.3872 | 0.3774 | 0.4907 | 0.4552 | 0.2646 | 0.3615 |
| IWAPU-1 | 0.4662 | 0.6495 | 0.7358 | 0.7484 | 0.7963 | 0.6949 | 0.6995 |
| IWAPU-2 | 0.4766 | 0.6563 | 0.7336 | 0.7451 | 0.7900 | 0.7021 | 0.7031 |
| IWAPU-3 | 0.4888 | 0.6368 | 0.7367 | 0.7428 | 0.7781 | 0.6879 | 0.6888 |
| IWAPU-4 | 0.4912 | 0.6407 | 0.7289 | 0.7385 | 0.7722 | 0.6849 | 0.6873 |
| IWAPU-5 | 0.5756 | 0.6392 | 0.6731 | 0.6842 | 0.7203 | 0.6325 | 0.6359 |
| IWAPU-6 | 0.5625 | 0.6360 | 0.7046 | 0.7050 | 0.7098 | 0.6225 | 0.6225 |
| IWAPU-7 | 0.5664 | 0.6174 | 0.6739 | 0.6765 | 0.6657 | 0.5878 | 0.6037 |
| IWAPU-8 | 0.5963 | 0.6231 | 0.6782 | 0.6830 | 0.7153 | 0.6030 | 0.6437 |
| SHZU-1 | 0.3564 | 0.4858 | 0.6023 | 0.6223 | 0.7104 | 0.4579 | 0.4612 |
| SHZU-2 | 0.0999 | 0.1115 | 0.5516 | 0.5617 | 0.4739 | -0.1625 | -0.0705 |
| SHZU-3 | 0.0846 | 0.0847 | 0.3936 | 0.5379 | 0.4348 | -0.1597 | 0.0000 |

**Table 8: Summary on the system descriptions for the spoken query set.**

| system ID | transcription type (target) | transcription type (query) | offline time [s] | index size [MB] | machine SPECH | online time [s] |
|---|---|---|---|---|---|---|
| BL-1 | K-REF-SYLLABLE-MATCH | K-REF-SYLLABLE-MATCH | 100 | 256 | Core i7 3720QM 2.6GHz, 16GB, no GPU | 30 |
| ALPS-1 | OWN | K-REF-WORD-MATCH | 3221 | 698 | Xeon E5-2630v3 2.4GHz, 128GB, GTX TitanX | 22975 |
| ALPS-2 | OWN | K-REF-WORD-MATCH | 3221 | 698 | Xeon E5-2630v3 2.4GHz, 128GB, GTX TitanX | 16603 |
| ALPS-3 | OWN | K-REF-WORD-MATCH | 223 | 630 | Xeon E5-2630v3 2.4GHz, 128GB, GTX TitanX | 16249 |
| ALPS-4 | OWN | K-REF-WORD-MATCH | 3221 | 698 | Xeon E5-2630v3 2.4GHz, 128GB, GTX TitanX | 19330 |
| ALPS-5 | OWN | K-REF-WORD-MATCH | 3221 | 698 | Xeon E5-2630v3 2.4GHz, 128GB, GTX TitanX | 22903 |
| ALPS-6 | OWN | K-REF-WORD-MATCH | 3221 | 68 | Xeon E5-2630v3 2.4GHz, 128GB, GTX TitanX | 8242 |
| IWAPU-1 | K-REF-WORD-MATCH | K-REF-WORD-MATCH | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 23373 |
| IWAPU-2 | K-REF-WORD-MATCH | K-REF-SYLLABLE-MATCH | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 21853 |
| IWAPU-3 | K-REF-WORD-MATCH | K-REF-WORD-MATCH | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 23060 |
| IWAPU-4 | K-REF-WORD-MATCH | K-REF-SYLLABLE-MATCH | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 21551 |
| IWAPU-5 | K-REF-WORD-MATCH | K-REF-WORD-MATCH | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 525 |
| IWAPU-6 | K-REF-WORD-MATCH | K-REF-WORD-MATCH | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 311 |
| IWAPU-7 | K-REF-SYLLABLE-MATCH | K-REF-SYLLABLE-MATCH | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 340 |
| IWAPU-8 | K-REF-SYLLABLE-MATCH | K-REF-SYLLABLE-MATCH | 30 | 50 | Core i7 4770 3.5GHz, 16GB, GTX 750Ti | 534 |
| SHZU-1 | K-REF-WORD-MATCH | K-REF-WORD-MATCH | 25292 | 5 | Xeon 2.4GHz, 32GB | 696 |
| SHZU-2 | K-REF-WORD-MATCH | K-REF-WORD-MATCH | 25292 | 5 | Xeon 2.4GHz, 32GB | 642 |
| SHZU-3 | K-REF-WORD-MATCH | K-REF-WORD-MATCH | 45220 | 1605 | Xeon 2.4GHz, 32GB | 4697 |